Special and General Relativity based on the Physical Meaning of the Spacetime Interval

Alan Macdonald
Department of Mathematics • Luther College
macdonal@luther.edu • http://faculty.luther.edu/~macdonal

Abstract

We outline a new and simple development of special and general relativity based on the *physical* meaning of the spacetime interval. The Lorentz transformation is not used.

1 Introduction

In 1908, only three years after Einstein's discovery of special relativity, Hermann Minkowski geometrized the theory by uniting space and time into spacetime. This provided enormous insight into special relativity and became the starting point for general relativity, a geometrization of gravity.

The fundamental geometric object in a Minkowski spacetime is the spacetime interval Δs between two events. Elementary special relativity texts discuss the interval to various degrees. In most, Δs is defined in terms of the coordinates of an inertial frame: $\Delta s^2 = \Delta t^2 - \Delta x^2$. In these texts the important properties of Δs are the formula just given, and its invariance under a change of inertial frames. The invariance is usually proved as a consequence of the Lorentz transformation. A physical meaning for Δs is given later, if at all.

My purpose here is to describe a new approach to relativity which puts the *physical meaning* of Δs front and center. I believe that this approach provides quicker access to and deeper understanding of both special and general relativity.

Elementary Euclidean geometry helps to motivate the approach. In Euclidean geometry, a distance Δs has a *physical* meaning as something *measured* by a ruler. Its meaning does not depend on the notion of coordinate systems. If coordinates are introduced, then we have the formula $\Delta s^2 = \Delta x^2 + \Delta y^2$ (Pythagorean theorem). But the *physical* meaning of Δs as a directly measureable quantity is more important than the *mathematical* formula for it in terms of coordinate differences (as important as the formula is). The formula for a rotation of coordinates is even less important. Note however, that since Δs has a physical meaning independent of coordinates, it is automatically invariant under a rotation of coordinates.

In Sec. 2 we discuss the interval of special relativity. The interval has a simple physical meaning as something *measured* by light, a clock, or a rod. It is thus a fundamental physical quantity. Its meaning does not depend on the notion of inertial frames. In the approach to special relativity described in Sec. 2, we define Δs physically, without reference to inertial frames. We then prove that if inertial frames are introduced, then $\Delta s^2 = \Delta t^2 - \Delta x^2$. But the *physical* meaning of Δs as a directly measureable quantity is more important than the *mathematical* formula for it in terms of coordinate differences (as important as the formula is). The Lorentz transformation is even less important. Note however, that since Δs has a physical meaning independent of coordinates, it is automatically invariant under a Lorentz transformation.

In Sec. 3 we develop the curved spacetime of general relativity by replacing the finite interval Δs with the infinitesimal interval ds and then repeating the development of special relativity in Sec. 2.

In Sec. 4, an appendix, we compare the development of general relativity presented here with those based on the equivalence principle.

2 The Interval in Special Relativity

The Hafele-Keating experiment provides the best introduction to the approach to special relativity taken here.[1] Recall that Hafele and Keating synchronized two clocks and then placed them in separate airplanes, which circled the Earth in opposite directions. When the clocks were brought together again, they were ticking at the same rate, but they were not synchronized. Both special and general relativistic effects contributed to the difference.[2] The experiment shows that clocks with different worldlines between two events in a spacetime can measure different times between the events, just as different curves between two points in a plane can have different lengths.

Two points in a plane determine a quantity Δs, the distance between the points as measured by a ruler. Analogously, we have the

Definition. Two events E and F in spacetime determine a physical quantity Δs, the *spacetime interval* between the events:

- If an inertial clock can move between E and F, define Δs to be the time between E and F as measured by the clock. Call Δs the *proper time* between the events and say that the events are *timelike separated*.
- If light can move between E and F, define $\Delta s = 0$. Say that the events are *lightlike separated*.
- If neither light nor clocks can move between E and F then, as we shall see, a rigid rod can have its ends simultaneously at E and F. (Simultaneously in the sense that light flashes emitted at E and F reach the center of the rod simultaneously, or equivalently, that E and F are simultaneous in the rest frame of the rod.) Define $|\Delta s|$ to be the rest length of this rod. (The reason for the absolute value will be clear later.) Call $|\Delta s|$ the *proper distance* between the events and say that the events are *spacelike separated*.

In each case, knowing Δs for a pair of events tells us something *physical* about the events. For example, if $\Delta s = 0$, then we know that light can move between the events.

Cartesian coordinates provide coordinates in a plane. Analogously, inertial frames provide coordinates in a flat spacetime. We use the notion of an inertial frame in the physical sense of Taylor and Wheeler[3]: a cubical lattice of inertial[4] rigid rods with synchronized clocks[5] at the nodes. Our fundamental assumption for special relativity is that the speed of light is the same in all inertial frames. Choose units of time and space so that $c = 1$.

The Pythagorean theorem allows us to compute the distance Δs between two points in terms of their coordinate differences: $\Delta s^2 = \Delta x^2 + \Delta y^2$. The following theorem, which may be considered the fundamental theorem of analytic Minkowskian geometry, allows us to compute the spacetime interval Δs between two events in terms of their coordinate differences.

Theorem. Let events E and F have coordinate differences $(\Delta t, \Delta x)$ in an inertial frame I. Then

$$\Delta s^2 = \Delta t^2 - \Delta x^2 \tag{1}$$

Proof. We prove Eq. 1 separately for E and F lightlike, timelike, and spacelike separated.

Lightlike. In this case light can move between the events. By definition, $\Delta s = 0$. Since $c = 1, \Delta x = \Delta t$. Eq. 1 follows.

Timelike. In this case an inertial clock C can move between the events. By definition, Δs is the time C measures between the events. Let C carry a rod R pointing perpendicular to its direction of motion. Let R have a mirror M on the end. At E a flash of light is sent along R from C. The length of R is arranged so that the flash is reflected by M back to F. Fig. 1 shows the path of the light in two spatial dimensions of I, together with C, R, and M as the light reflects off M.[6]

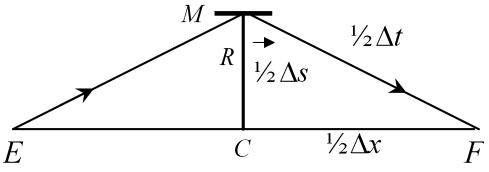

Figure 1: $\Delta s^2 = \Delta t^2 - \Delta x^2$ for timelike separated events E and F.

Refer to the rightmost triangle in Fig. 1. In I, the distance between E and F is Δx. This gives the labeling of the base of the triangle. In I, the light takes the time Δt from E to M to F. Since $c = 1$ in I, the light travels a distance Δt in I. This gives the labeling of the hypotenuse. C is at rest in some inertial frame I'. In I', the light travels the length of the rod twice in the proper time Δs between E and F measured by C. Since $c = 1$ in I', the length of the rod is $\frac{1}{2}\Delta s$ in I'. This gives the labeling of the altitude of the triangle. (We have tacitly assumed here that the length of the rod is the same in I and I'. Textbook authors make the same assumption when setting $y = y'$ and $z = z'$ for inertial frames whose x- and x'-axes coincide. This follows from a simple symmetry argument or the relativity principle.[7])

Applying the Pythagorean theorem to the triangle now shows, in a most graphic way, that accepting a universal light speed forces us to abandon a universal time and accept Eq. (1) for inertial clocks.

The argument shows that since the light travel distance is longer in I than for C (twice the hypotenuse vs. twice the altitude) and the speed $c = 1$ is the same in I as for C, the time (= distance/speed) between E and F is longer in I than for C.

Turning this around, the argument shows how it is possible for a single flash of light to have the same speed in inertial frames moving with respect to each other: the speed (distance/time) can be the same because the distance *and* the time are different in the two frames.

Spacelike. By definition, neither light nor clocks can move between spacelike separated events. Thus $\Delta x > \Delta t$. For convenience, take E to have coordinates $E(0,0)$. Fig. 2 shows the worldline W' of an inertial observer O' moving with velocity $v = \Delta t/\Delta x$ in I. (Note that Δt is the distance difference and Δx is the time difference.) $L\pm$ are the light worldlines through F. Since $c = 1$ in I, the slope of $L\pm$ is ± 1. Solving simultaneously the equations for $L-$ and W' gives the coordinates $R(\Delta x, \Delta t)$. Similarly, the equations for $L+$ and W' give $S(-\Delta x, -\Delta t)$.

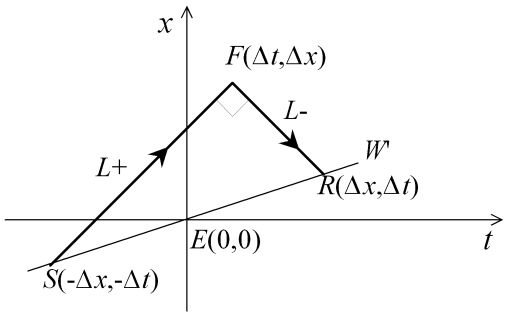

Figure 2: $\Delta s^2 = \Delta t^2 - \Delta x^2$ for spacelike separated events E and F. See the text.

According to Eq. (1), the proper time, as measured by O', between the timelike separated events S and E, and between E and R, is $(\Delta x^2 - \Delta t^2)^{\frac{1}{2}}$. Since the times are equal, E and F are, by definition, simultaneous in an inertial frame I' in which O' is at rest. Since $c = 1$ in I', the distance between E and F in I' is $(\Delta x^2 - \Delta t^2)^{\frac{1}{2}}$. Thus a rod at rest in I' with its ends simultaneously at E and F will have rest length $(\Delta x^2 - \Delta t^2)^{\frac{1}{2}}$. By definition, this length is $|\Delta s|$. This proves Eq. (1) for spacelike separated events, and completes the proof of the theorem.

For the timelike separated events above, $v = \Delta x/\Delta t$ is the speed in I of the inertial clock measuring the proper time Δs between the events. Then:

$$\Delta s = (\Delta t^2 - \Delta x^2)^{\frac{1}{2}} = \left[1 - (\Delta x/\Delta t)^2\right]^{\frac{1}{2}} \Delta t = (1 - v^2)^{\frac{1}{2}} \Delta t. \tag{2}$$

This is the time dilation formula, which expresses the proper time Δs between the events in terms of the inertial frame time Δt between the events.

For the spacelike separated events above, $v = \Delta t/\Delta x$ is the speed in I of the rod measuring the proper distance between the events. A calculation analogous to Eq. 2 shows that

$$|\Delta s| = (1 - v^2)^{\frac{1}{2}} |\Delta x|.$$

This expresses the proper distance $|\Delta s|$ between the events in terms of the inertial frame distance $|\Delta x|$ between the events. (This is *not* the length contraction formula.)

Time along worldlines. In a plane, the relation $\Delta s^2 = \Delta x^2 + \Delta y^2$ gives only the length of a special curve between two points: the straight line. But the differential version, $ds^2 = dx^2 + dy^2$, can be integrated to give the length of *any* curve between the points.

Eqs. (1) and (2) give only the time measured by a special clock moving between two events: an inertial clock. But the differential versions can be integrated to give the time s measured by *any* clock C moving between the events.[8] Thus let C move with velocity $v(t)$, $t_1 \le t \le t_2$. Then according to the differential version of Eq. (2),

$$s = \int_{t_1}^{t_2} ds = \int_{t_1}^{t_2} (1 - v^2(t))^{\frac{1}{2}} dt < \int_{t_1}^{t_2} dt = t_2 - t_1.$$

The time s measured by C between the events is less than the time $t_2 - t_1$ measured by the synchronized clocks of the inertial frame between the events. In particular, if C returns to its starting point P in the inertial frame, then it measures less time for the round trip than a clock at rest at P. This analysis of the "twin paradox" uses only one inertial frame.

Curvilinear coordinates (y^i) can be used in the flat spacetime of special relativity, just as they can in a plane. Such coordinates are not much used in flat spacetimes because inertial frame coordinates are usually easier to use. We will not have this luxury in curved spacetimes. As preparation for this, we express the differential version of Eq. (1) in terms of the (y^i). Write the differential version, with three space coordinates, as

$$ds^2 = f_{mn} dx^m dx^n,$$

where $(x^0, x^1, x^2, x^3) = (t, x, y, z)$ and

$$(f_{mn}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

(We use the Einstein summation convention, in which a repeated index is implicitly summed.) If the coordinate change is $x^m = x^m(y^j)$, then

$$\begin{aligned} ds^2 &= f_{mn} dx^m dx^n \\ &= f_{mn} \left(\frac{\partial x^m}{\partial y^j} dy^j \right) \left(\frac{\partial x^n}{\partial y^k} dy^k \right) \\ &= \left(f_{mn} \frac{\partial x^m}{\partial y^j} \frac{\partial x^n}{\partial y^k} \right) dy^j dy^k \\ &= g_{jk} dy^j dy^k, \end{aligned} \tag{3}$$

where we have set

$$g_{jk} = f_{mn} \frac{\partial x^m}{\partial y^j} \frac{\partial x^n}{\partial y^k}. \tag{4}$$

3 The Interval in General Relativity

The central idea of general relativity is that a spacetime containing mass is curved. This is not simply a negative statement that the spacetime is not the flat spacetime of special relativity. It is the positive statement that a spacetime containing mass has a metric ds of a special kind: a Riemannian metric, where $ds^2 = g_{jk}dy^j dy^k$. There are other possibilities. For example, we might have $ds^4 = g_{jklm}dy^j dy^k dy^l dy^m$ or, more generally, a *Finsler metric.*[9]

A curved surface is different from a flat surface. However, a simple observation of C. F. Gauss provides the key to the construction of the modern theory of surfaces: a small region of a curved surface is much like a small region of a flat surface. This is familiar: ignoring surface irregularities, a small region of the Earth appears flat.

Imagine *surface dwellers*, two dimensional beings inhabiting a flat surface. They define Δs as the distance between two points as measured by a rigid rod. They also construct coordinate systems of square grids of rods and discover that $\Delta s^2 = \Delta x^2 + \Delta y^2$. The surface dwellers awaken one morning to find that they can no longer fit rigid rods together to form a square grid over large areas. However, one of them, C. F. Gauss, notices that he can construct *small* square grids in which $ds^2 = dx^2 + dy^2$. He then translates, as in Eq. (3), this equation into a general coordinate system (y^1, y^2) to obtain a Riemannian metric $ds^2 = g_{jk}dy^j dy^k$. In this way he discovers that his universe became curved overnight.

A spacetime containing mass is different from a flat spacetime. For example, tidal effects prevent us from fitting inertial rigid rods together to form an inertial frame over all spacetime. However, a simple observation of Einstein provides the key to the construction of general relativity: *as viewed by inertial observers*, a small region of a spacetime containing matter is much like a small region of flat spacetime. We see this vividly in motion pictures of astronauts in orbit: inertial objects in their cabin move in a straight line at constant speed, just as in a flat spacetime.

In accord with Einstein's observation, in a spacetime containing mass it is possible to construct a *local inertial frame*: a small cubical lattice of inertial rigid rods with synchronized clocks at the nodes. We can imagine a local inertial frame in the astronauts' cabin. Our fundamental assumption for special relativity was that $c = 1$ in inertial frames. Again in accord with Einstein's observation, our fundamental assumption for general relativity is that $c = 1$ in local inertial frames.

We now repeat the development of special relativity in the last section for a spacetime containing mass:

1. *Define* the spacetime interval ds physically for neighboring events. The definition is given at the beginning of Sec. 2.

2. *Prove* that $ds^2 = dt^2 - dx^2$ in local inertial frames from our assumption that $c = 1$ in local inertial frames. This is the theorem following the definition.

3. *Transform* this equation to $ds^2 = g_{jk}dy^j dy^k$ in an arbitrary coordinate system. The transformation is given in Eqs. 3 and 4.

In this way we discover that a spacetime containing mass has a Riemannian metric! In short: *If $c = 1$ in local inertial frames, then spacetime has a Riemannian metric.*

In this approach to general relativity:

- The physical meaning of the interval precedes its mathematical expression.
- The physical and mathematical similarity between special and general relativity is emphasized.
- Tensor analysis is not needed.
- On the other hand, the calculation Eq. 3 shows that Eq. 4 holds between any two coordinate systems, i.e., the metric is a tensor. Thus the notion of a tensor arises naturally.
- The metric was constructed using local inertial frames. There is obviously a relationship between the motion of these free falling coordinate systems and the distribution of mass in the spacetime. Thus there is a relationship between the *metric* of a spacetime and the distribution of *mass* in it. This helps motivate the field equation $\mathbf{R} - \frac{1}{2}R\mathbf{g} = -8\pi\kappa\mathbf{T}$, which is of the form:

$$\left[\begin{array}{c}\text{quantity determined}\\ \text{by } \textit{metric}\end{array}\right] = \left[\begin{array}{c}\text{quantity determined}\\ \text{by } \textit{mass}\end{array}\right].$$

- The analogy with surface dwellers is helpful.

4 Appendix: The Equivalence Principle

The Riemannian metric of general relativity is usually taken to be a consequence of the equivalence principle.[10] However, Synge[11] and Ohanian[12] have argued, for me persuasively, that the equivalence principle, as usually understood, is false. I am not aware of any attempt to refute their arguments. I consider here the equivalence principle in the context of this paper.

It is again convenient to start with curved surfaces. Consider a triangle on a sphere. The sum of its angles exceeds 180°. But if the triangle's area is small, then the sum of its angles is approximately 180°: $\lim_{\text{Area}\to 0}(\text{Angle sum}) = 180°$. We say that the angle sum does not *couple to the curvature* of the sphere.

Surface dwellers might try to extend this by asserting an *equivalence principle*: for every measurement value M on a surface, $\lim_{\text{Area}\to 0} M = M_o$, where M_o is the flat surface value of M. However, this is false, as we now show. Let our sphere have radius R and let N be a point on it. Connect all points at distance r from N to get a circle C of radius r (in the sphere) and circumference $C(r)$. See Fig. 3. Let the measurement M on C be

$$M = \left(\frac{3}{\pi}\right) \frac{2\pi\, r - C(r)}{r^3}.$$

From the figure,

$$C(r) = 2\pi R \sin\phi = 2\pi R \sin(r/R) = 2\pi R \left[r/R - (r/R)^3/6 + \ldots\right].$$

Using this, we find that $\lim_{\text{Area}\to 0} M = 1/R^2$. Since $M_o = 0$, their equivalence principle is violated. The limiting value of M (on any surface) is called the *curvature* of the surface at N. The surface dweller's equivalence principle is violated because some local measurements couple to the curvature. Thus the most we can say is our statement above: "A small region of a curved surface is much like a small region of a flat surface."

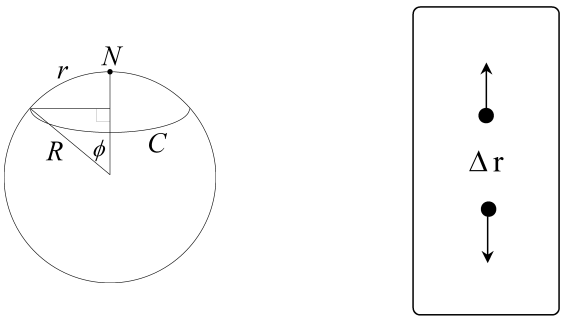

Figure 3: Violations of an equivalence principle.

Let us now turn to curved spacetimes. Consider with Einstein an elevator in radial free fall toward the Earth. Two particles are initially at rest in the elevator, one above the other, at a distance Δr apart. See Fig. 3. The acceleration

of each particle with respect to the Earth is $a = -\kappa M/r^2$. Thus their relative acceleration (which is measurable in the elevator) is $\Delta a \approx (2\kappa M/r^3)\Delta r$. Thus if $\Delta r \approx 0$, then $\Delta a \approx 0$: $\lim_{\Delta r \to 0} \Delta a = 0$.

The equivalence principle extends this by asserting: For *every* measurement value M in a region R of a curved spacetime, $\lim_{\text{size}(R) \to 0} M = M_0$, where M_0 is the flat spacetime value of M.[14] However, this statement is false. For suppose we measure, not Δa, but $\Delta a/\Delta r$. Then $\lim_{\text{size}(R) \to 0} M = da/dr = 2\kappa M/r^3$. Since $M_0 = 0$, *the equivalence principle is violated.*[15]

The $\Delta a/\Delta r$ measurement couples to the curvature of spacetime, whereas the Δr measurement does not. The equivalence principle is violated because some local measurements couple to the curvature. We cannot simply dismiss such measurements by saying that they "suffer tidal effects". For unless we can specify *physically* which measurements suffer such effects, this amounts to saying that "measurements obey the equivalence principle unless they do not". The most we can say is our statement above: "A small region of a curved spacetime is much like a small region of a flat spacetime."

Without the equivalence principle, how can we obtain the Riemannian metric? Ohanian simply postulates its existence.[16] Our approach is different: We have shown here that a *specific* physical principle, a universal light speed in local inertial frames, implies the existence of a Riemannian metric for curved spacetimes. Even if one believes the equivalence principle, this seems worth noting.

[1] J. C. Hafele and R.E. Keating, "Around the World with Atomic Clocks," Science **177**, 166-170 (1972).

[2] Consider a simple model of the experiment in which a clock circles a 40,000 km equator to the west at 1000 km/hr at a height of 10 km and another clock remains at rest on the ground. Special relativity predicts a difference of $1.4x10^7$ sec due to the clocks' velocities. The gravitational redshift predicts a difference of $1.6x10^7$ sec. The Schwartzschild metric predicts the sum of these differences.

[3] *Spacetime Physics*, E. F. Taylor and J. A. Wheeler, (Freeman, San Francisco, 1966), pp. 17-18.

[4] An object (in a gravitational field or not) is *inertial* if an accelerometer carried with it registers zero.

[5] For a careful discussion of clock synchronization, see A. L. Macdonald, "Clock Synchronization, a Universal Light Speed, and the Terrestrial Redshift Experiment," Am. J. Phys. 51, 795-797 (1983).

[6] Taylor and Wheeler (Ref. 3, p. 25) also use Fig. 1. But their purpose is different from ours: they define Δs by Eq. 1 and prove its invariance, whereas we define Δs physically and prove Eq. 1.

[7] E.F. Taylor and J.A. Wheeler, Ref. 3, pp. 21-22.

[8] *Introduction to Special Relativity,* W. Rindler, (Clarendon Press, Oxford, 1982), pp. 31-33. Rindler calls this the *clock hypothesis* and cites strong experimental evidence for it.

[9] See, for example, *Finsler Geometry, Relativity, and Gauge Theories,* G.S. Asanov,

(Reidel, Dordrecht, The Netherlands, 1985)

[10]C. Misner, K. Thorne, and J. Wheeler, *Gravitation* (Freeman, San Francisco, 1970) p. 207; W. Rindler, *Essential Relativity* (Springer-Verlag, New York, 1977), p. 114; S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972), p. 70.

[11]J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1971), p. ix.

[12]H. Ohanian, "What is the principle of Equivalence? ," Am. J. Phys. **45**, 903 (1977).

[13]For more information about the curvature of a surface and curvature in higher dimensions, see W. Rindler, *Essential relativity: special, general, and cosmological* (New York, Springer-Verlag, 1977), sections 7.1-7.3; and L. Eisenhart, *Riemannian geometry* (Princeton University Press, Princeton, 1926), sections 25 and 34.

[14]K. S. Thorne, D. L. Lee, and A. P. Lightman, "Foundations for a Theory of Gravitation Theories," Phys. Rev. D **7**, 3563 (1973). The authors give the most careful statement of the equivalence principle of which I am aware (p. 3572): "The outcome of any local, nongravitational, test experiment is independent of where and when in the universe it is performed, and independent of the velocity of the (free falling) apparatus." The authors carefully define "local, nongravitational, test experiment" on p. 3571. Among other things it requires "a sequence of measurements of successively smaller size ... until the experimental result approaches a constant value asymptotically." Their idea of a sequence of measurements recognizes that most measurements in a small elevator do not give exactly the flat spacetime value. The idea also avoids the notion of an infinitesimal region of spacetime. As convenient as it is to think about such regions, they do not exist, physically or mathematically.

[15]It appears to me that the da/dr measurement is a local, nongravitational, test experiment in the sense of Thorne, et. al. (Ref. 14.) Classically, da/dr can be measured in as small a region of space and time as desired. But the outcome of such a measurement is clearly dependent on where it is performed; it violates their equivalence principle.

[16]H. Ohanian, *Gravitation and Spacetime* (Norton, New York, 1976), p. 202, Eq. [21]